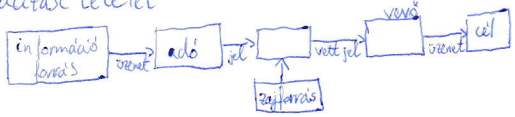
Információ és kódelmélet

Alapfogalmak:

- Kódok: - speciális jelentéssel bíró szavak, betűk, színek, jelek
  - nem minden kód titkos
  - a napi gyakorlatban a kódolást gyors és könnyű üzenetküldésre használjuk (szemafor, Morse, bináris kód...)
- Kriptológia: rejtett vagy titkos kommunikáció tudománya és felöleli mindazokat a módszereket, amelyekkel olyan üzeneteket lehet készíteni, amit csak az arra jogosultak tudnak megfejteni
- Kódáselmélet: olyan algoritmusok keresésének a tudománya, amelyekkel a digitális információt hatékonyan kódolhatjuk zajos csatornákon történő megbízható átvitelhez
- Információelmélet: az üzenetek adásának és vételének (kommunikációnak) matematikai elmélete

Történeti áttekintés:

- 1837: S. Morse elektromos távírója és kódja
- 1947-48: Az információelmélet születése: C. Shannon forráskódolási és csatornakapacitási tételei



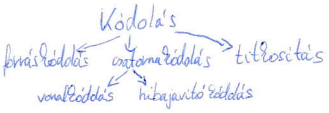
- 1949-50: Az első hibajavító kódok felfedezése (M. Golay, R. Hamming)
- 1962: Huffman kód
- Shannontól származó hírközlési rendszerek általános sémája:

tárolási rendszerek ált. sémája: forrás -> írókészülék -> memória -> olvasókészülék -> felhasználó

↳ pontosabban: forrás → tömörítés (forrás kódolás) → kódolás (csatorna kódolás) → zajos csatorna → dekódolás → felismerő

További fogalmak:

- Információ: véges számú, ismert lehetséges alternatíva valamelyikének megnevezése. Azt vizsgáljuk, hogy mennyi „információ" kell egy tetszőleges elem azonosításához, ha az alternatívák lehetséges A halmaza adott.
- Forrás: az információt (közleményeket) szolgáltató objektum
- Forrásábécé: véges jel halmaz (ábécé)
- Közlemény: a „forrásábécé" jeleiből álló véges jelsorozat (tkp. az információ)
- Távközlési csatorna: egyenlő időközönként egymásután következő jelek (csatorna kódjelek) juthatók át rajta
- Csatornaábécé: a csatornán továbbítható kódjelek összessége
  - zajmentes csatorna: ideális, a jeleket nem torzítja, a kimeneti kódjel mindig ugyanaz mint a bemeneti
  - zajos csatorna: a jeleket torzíthatja
  - bináris csatorna: kétféle kódjel továbbítható rajta
- Kódoló: a közleményt átalakítja a csatornán való továbbításhoz (kódolás)
- Kódközlemény: az eredeti közlemény kódolt alakja, a csatornaábécé „betűiből" álló véges sorozatok
- Dekódoló: a csatorna kimeneti oldalán vett (kód)közlemény megfejtése (dekódolás)
- Kódolási eljárás: olyan „utasítás", amely minden lehetséges közleményhez hozzárendel egy kódjelekből álló sorozatot, az illető közlemény kódközleményét
- Lehetséges kódolási eljárások:
  - betűnként, blokkonként
  - információ veszteség nélkül, információ veszteséggel
- Lehetséges kódtípusok: változó hosszúságú, fix hosszúságú (blokk) kód



-> Def: Egy kód blokk kód, ha a kódoló a csatornaábécé fix hosszúságú sorozataiból állnak.

A^+ -> az összes hosszúságú rendezett n-es (1 hosszú, 2 hosszú,... szavak) valamilyen halmazból Descartes vagy direkt szorzata

Információelmélet alapfogalmak, jelölések

Def: Tetszőleges véges $\Sigma \neq \emptyset$ halmazt ábécének nevezünk. A Σ ábécé elemeit a Σ betűinek (szimbólumainak) nevezzük.

pl: $\Sigma_{bool} = \{0,1\}$ a Boole ábécé

$\Sigma_{latin} = \{a, b, c, \dots z\}$ a latin ábécé

Def: A Σ ábécé jeleinek tetszőleges véges sorozatát Σ feletti szónak nevezzük. A w szó $|w|$ hossza a w-ben levő jelek száma.

A $w = x_1 x_2 \dots x_n$ szó hossza: $|w| = n$. A w szót felfoghatjuk a Σ^n halmaz egy $(x_1, x_2, \dots x_n)$ elemének is, amelyekből a zárójeleket és az elválasztójeleket elhagyjuk.

Def: Σ^+ a Σ ábécéből képezett összes szó halmaza

A továbbiakban jelöljön $\Sigma_x = \{x_1, x_2, \dots x_a\}$ egy forrásábécét, $\Sigma_y = \{y_1, y_2, \dots y_r\}_{r=2}^{\{0,1\}}$ pedig egy csatornaábécét.

Def: (betűnkénti, vagy szimbólum kód): A $K: \Sigma_x \to \Sigma_y^+$ $_{k=26}$ leképezést kódnak nevezzük, $K(x)$ az $x \in \Sigma_x$-hez tartozó kódszó, $\ell(x)$ a $K(x)$ kódszó hossza, $L_i = \ell(x_i)$

Kiterjesztett K^+ kód: $K^+: \Sigma_x^+ \to \Sigma_y^+$ leképezés

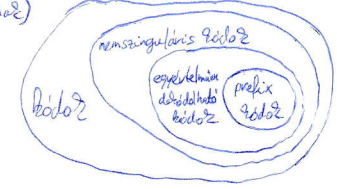
Def: Egy $K: \Sigma_x \to \Sigma_y^+$ kódot nemszingulárisnak nevezünk, ha minden $x, x' \in \Sigma_x$, $x \neq x'$ esetén $K(x) \neq K(x')$ is teljesül. (=> szingularis: van olyan ábécé betűje, amihez ugyanazt rendeljük)

A nemszingularitás elég egy kódjel egyértelmű azonosításához, de nem elég egy üzenet azonosításához.

Def: Egy $K: \Sigma_x \to \Sigma_y^+$ kód egyértelműen dekódolható, ha Σ_x^+ különböző elemeit (üzeneteket) különböző szavakat sorozatokba képzi, azaz: $\forall x, y \in \Sigma_x^+$, $x \neq y \Rightarrow K^+(x) \neq K^+(y)$

! Az egyértelmű dekódolhatóság több mint a K leképezés invertálhatósága, K^+ leképzést kell tudnunk invertálni. $K_{inv}^+: \Sigma_y^+ \to \Sigma_x^+$ minden $x \in \Sigma_x^+$ esetén $K_{inv}^+(K(x)) = x$

Def.: egy kód prefix, ha egyetlen kódszó sem kezdete (prefixe) egy másiknak (egyik kódszó sem folytatása egy másiknak)



Tétel: minden prefix kód egyértelműen dekódolható

→ Az állandó kódhosszúságú kód mindig prefix, ha a kódszavai különbözők

Gazdaságosság:

↳ egy üzenet továbbítása a lehető legrövidebb ideig vegye igénybe a csatornát

1, Állandó hosszúságú betűnkénti kódolás

L: kódjelek hossza

N jel hosszúságú üzenet esetén teljes cost = N·L, egy betű átlag költsége: $C = \frac{1}{N} \cdot N \cdot L = L$

2^L db L hosszú 0-1 sorozat van, egyértelmű dekódolhatóság akkor lehetséges, ha $d \le 2^L \rightarrow \log_2 d \le L$

Tehát $L = \lceil \log_2 d \rceil$ amikor is $\log_2 d \le C < \log_2 d + 1$

$\log_a b = \frac{\log_c b}{\log_c a}$

2, Változó hosszúságú betűnkénti kódolás

el. (x_i)	...	x_d
gyakoriság (N_i)	...	N_d
előfordulási valószínűség (P_i)	...	$P_d = \frac{N_d}{N}$
kódszó (K_i)	...	K_d
kód hossza (L_i)	...	L_d

$$\bar{L} = \frac{L}{N} = \sum_{i=1}^{d} \frac{N_i}{N} L_i = L_1 P_1 + L_2 P_2 + \dots L_d P_d = \sum_{i=1}^{d} L_i P_i$$ → egy betűre eső átlagos kódhossz

→ Ilyen esetben azt a kódolási eljárást tartjuk jobbnak, amelynél az L(K) átlagos kódhossz kisebb

Információ-entrópia

Információ: valaki véges számú, előre ismert esemény közül annak megnevezése, hogy melyik következett be. Alternatív megfogalmazás: az információ mértéke azonos azzal a bizonytalansággal, amelyet megszüntet.

Hartley: m számú azonos valószínűségű esemény közül egy megnevezésével nyert információ $I = \log_2 m = -\log_2 \frac{1}{m}$ (log m bázissal azonosítható egy elem) $I = -\log_2 P_i$

Shannon: minél váratlanabb egy esemény, bekövetkezése annál több információt jelent, annál több bizonytalanságot kell kiküszöbölni.

Legyen $A = \{A_1, A_2 \dots A_m\}$ esemény-halmaz, az A_1 esemény valószínűsége P_1, az A_m-é P_m. Ekkor A_i megnevezésével nyert információ: $I_{(A_i)} = \log_2 \frac{1}{P_i} = -\log_2 P_i$

$I_{(A_i)}$ → 1. Csak az esemény valószínűségétől függ(het)ve

2. nem negatív $I \ge 0$

3. Additív: ha $m = m_1 \cdot m_2$, $I_{(m_1 \cdot m_2)} = I_{(m_1)} + I_{(m_2)}$

Entrópia: Legyen X valószínűségi változó $\Sigma_X = \{x_1, x_2 \dots x_n\}$ az X lehetséges kimeneteinek a halmaza, $P_i = P(X = x_i) \ge 0$ pedig az x_i kimeneti valószínűsége. A $P_X = \{P_1, P_2 \dots P_n\}$ halmaz az X valószínűségeloszlását jelöli ($\sum_{i=1}^{n} P_i = 1$)

Az X valószínűségi változó entrópiáját a $H(X) = \sum_{i=1}^{n} P_i \cdot \log_2 \frac{1}{P_i} = -\sum_{i=1}^{n} P_i \log_2 P_i$ képlettel definiáljuk. ($I_{(x_i)}$) $H(X) \ge 0$, $H(X) \le \log_2 n$

$H(X) = \sum_{i=1}^{n} I_{(X=x_i)} \cdot P_{(X=x_i)}$ → Tehát az entrópia az X egyes értékei által tartalmazott információ mennyiségének várható értéke. Az entrópiát ezért az információ átlagos mennyiségének is nevezzük.

Def: A K szimbólum kód átlagos (várható) szóhossza: $L(K) = \sum_{x \in \Sigma_X} P_{(x)} \cdot l_{(x)}$ → $\frac{H(x)}{\log_2 r} \le L(K) < \frac{H(x)}{\log_2 r} + 1$

Kódfa

Minden kódszónak egy, a kezdőpontból kiinduló törött vonal felel meg, melynek egyenes szakaszait éleknek, töréspontjait szögpontoknak, a törtvonalakat a kódszóhoz tartozó ágnak nevezzük. $K = \{0, 100, 1010, 1011, 110, 111\}$ kódfája:

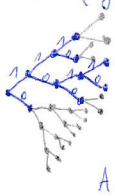
Prefix kód esetén minden kódszónak pontosan egy ág felel meg, egyik kódszóhoz tartozó ág sem lehet egy másik kódszó ágának folytatása.

(r->2 kódjel esetén szögpontokból legfeljebb r él indulhat ki)

Tétel: (Kraft-Fano egyenlőtlenség): Ha $K = \{K_1, K_2, \dots K_d\}$ r számú kódjelből készített prefix kód, ahol a K_i kódszó hossza L_i, akkor teljesül a

$r^{-L_1} + r^{-L_2} + \dots r^{-L_d} \le 1$ egyenlőtlenség.

Bizonyítás: Legyen $m = \max_i L_i$ a leghosszabb kódszó hossza. A kód minden szavát (kódfájának minden ágát) egészítsük ki mindenütt m hosszúságú kódszóra (ágakra) az összes lehetséges módon. Az előző példa alapján, ahol m=4



A vastag(ra ír)ék ekkor az eredeti ágakat jelölik. Az eredeti kódfa egy L_i hosszúságú ágát r^{m-L_i} féleképpen lehet kiegészíteni m hosszúságúvá a hiányzó $m-L_i$ él berajzolásával. A kiegészítéssel a kódfának összesen $r^{m-L_1} + r^{m-L_2} \dots r^{m-L_d}$ ága lesz. Az olyan kódfának, amelynek minden ága m hosszúságú és minden szögpontból r él indul ki, éppen r^m ága van. Ezért fennáll, hogy $r^{m-L_1} + r^{m-L_2} + \dots + r^{m-L_d} \le r^m$, ahonnan r^m-el való osztással a Kraft-Fano egyenlőtlenséget kapjuk.

A Kraft-Fano egyenlőtlenség szükséges feltételt ad. A bizonyítás alapján sejthető, hogy az egyenlőtlenség teljesülése esetén tudunk prefix kódot konstruálni.

A Kraft-Fano egyenlőtlenség fordítva is bizonyítható, ami az elégséges feltételt adja. A McMillan tétel pedig kimondja, hogy ez nem csak prefix kódokra igaz, hanem minden egyértelműen dekódolható kódra.

Prefix kód alsó határa: $\left(\frac{L(K) = \sum_{i=1}^{d} L_i P_i \to \min}{r^{-L_1} + r^{-L_2} \dots r^{-L_d} \le 1}\right\}$ feltételes szélsőértékfeladat megoldása

Tétel (Shannon): Legyen a $\Sigma_X = \{x_1, x_2 \dots x_d\}$ forrásábécé eloszlása $P_X = \{P_1, P_2 \dots P_d\}$ és legyen ennek $K = \{K_1, K_2 \dots K_d\}$ az $\Sigma_Y = \{y_1, y_2 \dots y_r\}$ csatornaábécé jeleiből allokált prefix kódja, ahol a K_i kódszó hossza L_i. Ekkor $L(K) = \sum_{i=1}^{d} L_i P_i \ge \frac{H(x)}{\log_2 r}$, ahol egyenlőség csak akkor állhat fenn, ha $P_i = r^{-L_i}$ minden i-re.

Bizonyítás nem kell csak, hogy honnan jött a megállapítás: egy szélsőérték problémából/optimalizációs problémából eredt (feltételt és célfüggvényt tudni)

Felső határ: Tétel: Bármilyen is legyen az Σ_X forrásábécé P_X eloszlása, a r elemű csatornaábécé jeleiből mindig készíthetünk olyan prefix kódot, amelyre:

$L(K) < \frac{H(x)}{\log_2 r} + 1$

} átlagos kódhossz határai = forráskódolási tétel

Hatásfok: $\eta = \frac{H(X)}{L(K)\log_2 r}$

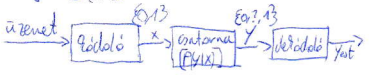
A maximális hatásfokú egyértelműen dekódolható kódszót optimális kódszónak nevezzük. Ilyen több is lehet, de biztosan van közöttük prefix (irreducibilis) kód.

Huffman-kód: optimális prefix kódot ad

Shannon-Fano:
1. valószínűségek meghatározása
2. csökkenő sorrendbe rendezés
3. elemek bevájzolása a $(0;1]$ intervallumon
4. az intervallum kiszegetése, amíg már csak 1 szimbólum marad benne
Határon lévő szimbólum esetén, a jobb oldali intervallumba soroljuk

Gilbert-Moore:
ugyanaz mint a Shannon-Fano, csak nem rendezzük csökkenő sorrendbe

A távközlési csatorna: egy olyan rendszer, amelynek bemenete véletlenszerűen függ a csatorna inputjától. A csatornát az átviteli valószínűségek $[P(y|x)]$ mátrixával jellemezzük, ahol x a csatorna input jele, y pedig a csatorna output jele.

zajos távközlési csatorna: 

A diszkrét emlékezet nélküli csatorna minden időegység alatt veszi az $x_1, \dots x_k$ jelek valamelyikét és reakcióként kibocsátja az $y_1 \dots y_n$ jelek valamelyikét. A csatorna azért „diszkrét" mert csak véges sok input és output jele van. Az „emlékezet nélküliség" azt jelenti, hogy az aktuális kimenet csak az aktuális inputtól függ, de nem függ a megelőző jelektől és továbbításuk eredményességétől.

$P_{(X|Y)} = \frac{P_{(X,Y)}}{P_{(Y)}}$

$P_{(Y|X)} = \frac{P_{(X,Y)}}{P_{(X)}}$

A továbbiakban $P_{(Y|X)}$ jelöli annak a valószínűségét, hogy egy x inputra y kimenetet kaptunk

$P_{(X,Y)}$	Y_1	Y_2	Y_3	
X_1	0,3	0,1	0,15	0,55
X_2	0,05	0,34	0,06	0,45
	0,35	0,44	0,21	1

(együttes valószínűségi mátrix)

$P_{(Y\|X)}$	Y_1	Y_2	Y_3	
X_1	0,55	0,18	0,27	1
X_2	0,11	0,76	0,13	1

csatornamátrix (átviteli mátrix)

Feltételes entrópia

$H(X|Y) = \sum_{y \in \Sigma_Y} P(y) \cdot H(X|Y=y) = \sum_{x \in \Sigma_X} \sum_{y \in \Sigma_Y} P_{(x,y)} \log_2 \frac{1}{P_{(x|y)}}$ → X valószínűségi változó feltételes entrópiája Y feltétel mellett

$H(X|Y) \geq 0$

$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$

$H(X|Y) \leq H(X)$

A $H(X)$ entrópia méri az X-re vonatkozó bizonytalanságot Y ismerete előtt, $H(X|Y)$ pedig az Y ismerete után. Ezért $H(X) - H(X|Y)$ különbség azt az információt mutatja, amelyet Y ismerete nyújt X-ről.

Def.: Az X és Y valószínűségi változó kölcsönös információjának nevezzük az $I(X,Y) = H(X) - H(X|Y) = \sum_{x \in \Sigma_X} \sum_{y \in \Sigma_Y} P_{(x,y)} \log_2 \frac{P_{(x,y)}}{P_{(x)} P_{(y)}}$ mennyiséget.

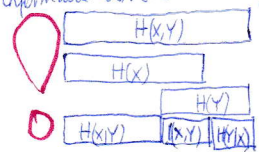
A kölcsönös információ tehát annak az információnak az átlagos mennyisége, amelyet X és Y értékei egymásra vonatkozóan tartalmaznak.

$I(X,X) = H(X)$

$I(X,Y) = I(Y,X)$

$0 \leq I(X,Y) = I(Y,X) \leq \min\{H(X), H(Y)\}$

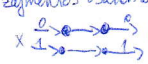
$H(X,Y)$ teljes információ tartalmának Venn-féle felbontása:



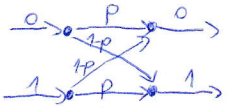
$H(X,Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \cdot \log_2 r_{ij}$

Csatornák és kapacitásuk

Def.: zajmentes csatorna, ha $X = Y$. Ilyenkor $I(X,Y) = I(X,X) = H(X) \leq \log r$



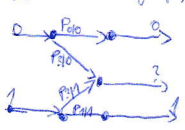
átmenet valószínűségei: $P_{0|0} = 1$ $P_{0|1} = 0$ $P_{1|0} = 0$ $P_{1|1} = 1$



$P_{0|0} = P$ $P_{0|1} = 1-P$ $P_{1|0} = 1-P$ $P_{1|1} = P$

Def.: bináris szimmetrikus csatorna: bármelyik kódjel helyes továbbításának valószínűsége p, hibás továbbításának valószínűsége 1-p

Def.: bináris törléses csatorna: bemenet $\{0,1\}$, kimenet $\{0,?,1\}$

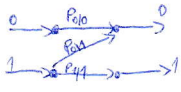


$P_{0|0} = 1-q$ $P_{0|1} = 0$

$P_{?|0} = q$ $P_{?|1} = q$

$P_{1|0} = 0$ $P_{1|1} = 1-q$

Def.: Z csatorna: $\{0;1\}$ bemenet, $\{0;1\}$ kimenet



$P_{0|0} = 1$ $P_{0|1} = q$

$P_{1|0} = 0$ $P_{1|1} = 1-q$